Racial Discrimination in Social Media Customer Service: Evidence from a Popular Microblogging Platform

Priyanga Gunarathne
University of Pittsburgh
priyanga.gunarathne@katz.pitt.edu

Huaxia Rui
University of Rochester
huaxia.rui@simon.rochester.edu

Abraham Seidmann
University of Rochester
avi.seidmann@simon.rochester.edu

Abstract

The concept of racial inequality has existed from the early days of service provision, with evidence dating back to ancient civilizations. While the emergence of the Internet and social media has drastically transformed almost every aspect of everyday life, including the intrinsic values of social relationships, the impact of racial disparities on receiving services on online platforms is not so evident. Although many consumer brands provide customer service on social media today, little is known regarding the prevalence and magnitude of racial discrimination in the context of social media customer service. Thus, in this study, we examine the existence and the extent of racial discrimination against African-Americans in social media customer service. We analyzed all complaints to seven major U.S. airlines on Twitter for a period of nine months. Interestingly, our empirical analysis finds that African-American customers are less likely to receive brand responses to their complaints on social media. To the best of our knowledge, this is the first study to empirically analyze the racial discrimination phenomenon in the context of social media customer service.

Keywords: African-American, customer service, deep learning, racial discrimination, social media, Twitter

1. Introduction

The concept of racial inequality has existed from the early days of service provision, with evidence dating back to ancient civilizations. Recent incidents of alleged discrimination against airline passengers of ethnic minorities have created a lot of bad press and public relations debacles for some major airlines in the U.S. Over the past fifty years, considerable societal efforts are made to reduce the level of discrimination against African-Americans in the United States [1]. Although much has improved lately, racial profiling and ethnic discrimination are still commonplace in many communities.

Marketplace discrimination has long been studied across a multitude of consumption contexts including the labor market, the housing market, short-term rental market, and car dealerships. A rich terminology is frequently used in the literature and popular press to refer to this phenomenon – for example, consumer racial profiling (CRP) [2], consumer discrimination [3], and shopping while black (SWB) [4]. It refers to the differential treatment that the consumers of a minority group (e.g., African-Americans, Muslims, women, immigrants, etc.) receive (usually a denial or a degradation of products and/or services), compared to the consumers of a majority group with otherwise identical characteristics [5].

With the emergence of the Internet, markets changed dramatically and a considerable portion of transactions has moved online. As online markets can be more anonymous than in-person transactions, there may be less room for racial discrimination in the online marketplace [1]. For example, Ayres and Siegelman [6] find that African-American car buyers are charged higher prices than white car buyers at dealerships, although Morton et al. [7] find no such difference amongst races in online purchases. Likewise, large online retailers such as Amazon and eBay offer little scope for discrimination, as sellers effectively pre-commit to accept all buyers regardless of race or ethnicity [1]. Nonetheless, the rise of social media has both fueled and been fueled by an unprecedented amount of public sharing of personal information that frequently includes disclosures and revelations of individuals, sometimes in surprisingly candid nature [8]. It has enabled people to publicly reveal multiple facets of themselves, including their private lives, social lives, and opinions [9], perhaps creating a new channel of discrimination along a multitude of social dimensions.

With recent advancements of social media and smart phones, consumers have increasingly become mobile and it is very easy and convenient to interact with brands online, rather than contacting the call

URI: https://hdl.handle.net/10125/60116
ISBN: 978-0-9981331-2-6

center and waiting to connect with a live-agent. In response to such requests for engagement, many consumer brands provide customer service on social media today, extending their main-stream customer service to the social space.

The current scholarship on racial discrimination in customer service has mostly examined the phenomenon in traditional contexts and little is known regarding the prevalence and magnitude of racial discrimination in the context of social media customer service. Thus, in this study, we investigate the existence and the extent of racial discrimination in the context of social media customer service. Specifically, we focus on the following research question: *Are brands less likely to respond to a complaint sent to it by an African-American customer, in the context of social media customer service?* To address this research question, we select Twitter as the social media platform and focus on the airline industry, as airlines have extensively leveraged Twitter for real-time customer service. We analyzed all tweets mentioning the official Twitter accounts of seven major U.S. airlines for a period of nine months, using text-mining techniques to extract and process data in a scalable fashion. Our empirical analysis suggests that African-American customers are less likely to receive brand responses to their complaints on social media, suggesting potential racial discrimination in social media customer service.

The rest of the paper is organized as follows. We first review relevant literature and then develop the hypotheses for our research questions. After describing our data and measures, we present our deep learning model for user classification. Then we estimate our main econometric model and present the empirical results. We conclude the paper by discussing the implications of our findings.

2. Literature Review

Marketplace discrimination has received a great deal of attention over the past few decades and has been the focus of many studies in economics and marketing. Edelman et al. [1] investigate the existence and extent of racial discrimination against ethnic minority guests on Airbnb, a popular online marketplace for short-term rentals. In a field experiment on Airbnb, they find that applications from guests with distinctively African-American names are 16 percent less likely to be accepted relative to identical guests with distinctively white names. They further find that discrimination occurs among landlords of all sizes, including small landlords sharing the property and larger landlords with multiple properties.

Ge et al. [10] examine the opportunity, peer transportation companies such as Uber and Lyft present, to rectify long-standing discrimination or worsen it. They sent passengers in Seattle, WA and Boston, MA on nearly 1,500 rides on controlled routes and recorded key performance metrics. Their findings indicate a pattern of discrimination, which they observed in Seattle through longer waiting times for African American passengers—as much as a 35 percent increase. In Boston, they observed discrimination by Uber drivers via more frequent cancellations against passengers when they used African American sounding names. Across all trips, the cancellation rate for African American sounding names was more than twice as frequent compared to white sounding names.

Laouénan and Rathelot [11] also use data from Airbnb collected in 19 major cities in North America and Europe to measure discrimination against ethnic-minority hosts. Using the ratings that provide potential guests with information about the quality of a listing, they build a credible measure of statistical discrimination. They find that hosts from a minority ethnic group charge 16 percent less than other hosts in the same cities and an additional review increases the daily price more for minority than for majority hosts. Estimating the parameters of a theoretical pricing model, they further find that statistical discrimination accounts for most of the price differential.

Acquisti and Fong [8] examined hiring discrimination in the U.S. labor market, focusing on personal information posted by job candidates on social media sites. They created profiles for job candidates on popular social networks and submitted job applications on their behalf to over 4,000 employers. After comparing interview invitations for a Muslim versus a Christian candidate, and a gay versus a straight candidate, they find no difference in callback rates for the gay candidate compared to the straight candidate, but a 13% lower callback rate for the Muslim candidate compared to the Christian candidate. Their results suggest that the online disclosure of certain personal traits can influence the hiring decisions of U.S. firms and the likelihood of hiring discrimination via online searches may vary across employers.

Bartoš et al. [12] monitor information acquisition in field experiments on discrimination and examine whether gaps arise when decision makers choose the effort level for reading an application. In both countries they study, negatively stereotyped minority names are shown to reduce employers' effort to inspect resumes. Furthermore, they find minority names to increase information acquisition in the rental housing market.

All these studies have provided important insights on racial discrimination mostly in the context of housing and labor markets. However, the studies that examined the phenomenon in customer service, particularly in the context of social media are almost non-existent. Our paper fills this gap and contributes to the stream of literature on racial discrimination in social media customer service.

3. Development of the Hypothesis

At the turn of the twentieth century, in the early years of the economics profession in the United States, many leading economists argued for Nordic superiority and Black inferiority [13]. Before 1960s, economists rarely concerned themselves about the problems of racial discrimination and racial inequality in society [14]. Since Gary Becker's [15] *The Economics of Discrimination* that provided the first sustained theoretical treatment of the subject, economists have been deeply interested in the presence of discrimination in the marketplace [16]. Thus, two workhorse models of discrimination have been developed in the economic literature.

In the first model [15] developed for the context of the labor market, some employers have a distaste for hiring members of the minority group. This distaste may lead them to refuse hiring members from the minority group, or if they do hire them, pay them less than the other workers for the same level of productivity. Thus, if the conditions of perfect competition are satisfied, discriminating employers will experience low profits and would be wiped away and taste-based discrimination would disappear. Hence, taste-based discrimination is considered clearly inefficient.

The second model deals with statistical discrimination [17, 18] that views the differential treatment to the members of the minority group as a consequence of imperfect information, and the discrimination as the result of a signal extraction problem. A classic example is an employer assessing the expected productivity of a worker he is considering hiring [16]. The employer is privy to some imperfect signal of productivity, such as the impression he gets in an interview, or by reading the résumé of the applicant. He is also aware of the job applicant's group membership such as his race and gender. More informative and complete the signal of the individual applicant is, the greater will be the weight placed on that information. On the other hand, when the person specific information is limited, the employer might evaluate the group-specific membership to obtain additional valuable information of expected productivity. For example, if it is known to the employer that minority applicants are on average less productive than majority applicants, and he sees two applicants with similar and unbiased signals of productivity, he should rationally favor the majority applicant to the minority one as her expected productivity is higher. Thus, statistical discrimination will result in some minority-group workers being treated less favorably than majority-group workers of the same level of true productivity.

Both taste-based discrimination and statistical discrimination could be the driving forces of discrimination in the context of social media customer service. For example, the innate qualities of certain social media agents may make them dislike African-American individuals, preventing them from associating positively with African-American customers. On the other hand, in cases where complete information regarding the perceived value of a customer is not available, certain social media agents may choose not to invest their time to search for missing pieces of information, but rather choose to misinterpret the value of customers based on their racial identities and consequently to let their complaints go nowhere.

Although several studies have been successful in documenting evidence that discrimination exists, many found it difficult to link the patterns of discrimination to any specific theory. Blurring the sharp line economists tend to draw between taste-based and statistical explanations of discrimination, psychologists have made a considerable progress in advancing theories and conducting laboratory experiments that have been helpful in better understanding the micro-foundations of discrimination [19].

Studies on discrimination in psychology are often related to the concept of prejudice, usually characterized as an unjustified or incorrect attitude (usually negative) towards an individual, based on that individual's membership of a social group. At least two dominant views of prejudice in psychology literature could serve as foundations to the previously discussed animus-based models of discriminatory behavior from economics. First, prejudice could make group membership an important component of social identity. Seminal work by Tajfel [20] and Tajfel and Turner [21] provide experimental evidence that demonstrates the major role social identity plays in the underlying process of prejudice. They show that the mere assignment of individuals into groups (even totally arbitrary ones that do not last, and with no objective conflict of interest or hostility between groups), is sufficient to produce favoritism for in-group members and negative attitudes toward out-group members.

Second, prejudice could trigger unconscious hence unintentional forms of discriminatory behavior of

individuals. In other words, it is believed that attitudes can occur in implicit modes and people can behave in ways that are inconsistent or sometimes even opposite to their explicit views of self-interests [22, 23, 24]. Neurophysiologists have shown that race perception influences and regulates cognitive processes such as affection [25, 26] and stereotype [27, 28]. Furthermore, studies in neuroscience show that different regions of the brain are activated in conscious versus unconscious processing, suggesting that unconscious processes are unique mental activities. For example, unconscious processing of a black face is associated with activations of area of the brain related to emotions and fear, while the conscious processing of the same face activates the areas of the brain associated with control and regulation [19]. This kind of implicit or unintentional biases could occur under conditions of high time pressures, cognitive loads, and ambiguity.

Hence, the conscious or unconscious prejudice against the minority-group members could imply that the individuals could be ignorant about the quality of minorities, triggering more statistical discrimination [29]. Therefore, in the context of social media customer service, if the social need to positively associate with the majority-group members also makes the minority-group members feel more distant and unknowable, social media agents may choose not to invest in resolving complaints of a minority-group member, or decide that the individual signals of perceived value (e.g., honesty, trustworthiness, and genuineness) of minority-group customers are totally uninformative. Building upon the theories of discrimination from economics and psychology, we propose the following hypothesis for empirical testing.

H1: An airline is less likely to respond to a complaint sent to it by an African American customer, relative to non-African American customers.

4. Data, Measures, and Methodology

Earlier research focused on three main techniques to measure discrimination in various consumption contexts: regression analysis, audit studies, and correspondence studies. We briefly describe each of these techniques next.

The regression methodology usually employs some consumption outcome (e.g., price) as the dependent variable, and group membership indicators along with relevant controls, as the explanatory variables [30]. The test for discrimination is whether the coefficient of group membership variable is significant. In audit studies, two individuals (i.e., auditors) are matched for all relevant personal characteristics other than the one that is presumed to lead to discrimination (e.g., race, gender) and then they apply for a job, a housing unit, or a mortgage, or begin to negotiate for a good or a service. The results they achieve and the treatment they receive in the transaction are closely monitored to determine if the outcomes reveal patterns of discrimination based on the trait studied [31]. Unlike audit studies that rely on real auditors that meet with a potential employer or a landlord, correspondence studies rely on fictitious applicants [19]. For instance, in response to a job or rental advertisement, the researcher sends several pairs of résumés, letters or emails of interest, one of which is assigned the perceived minority trait. Discrimination is assessed by comparing the outcomes (e.g., callbacks from employers or landlords) for the fictitious applicants with and without the particular minority trait in question.

In this study, we leverage the abundance of data generated from the actual interactions between customers and brands (i.e., companies) on social media in real-time, to reveal any patterns of race-based differential treatment. Specifically, we focus on how major airlines in the U.S. respond to customer complaints on Twitter. We collected all tweets mentioning the official Twitter accounts of seven major U.S. airlines in the 2014 – 2015 time period.

To distinguish complaints (i.e., tweets) from all other types of tweets, we followed a lexicon-based approach to build a complaint classifier. Based on our reading of a few thousand random tweets sent by customers to airlines, we developed two lexicons that contain n-grams of complaint related key-words (negative lexicon) and compliment related key-words (positive lexicon) respectively. A customer tweet was categorized as a complaint, if it matched at least one term in the negative lexicon and none in the compliment lexicon. We report an 84.5% precision for our lexicon-based complaint classifier.

On twitter, interactions between users usually flow through conversations. For example, a complaint started as a single tweet, may continue along a series of tweets exchanged between the customer and the airline forming a conversation. Thus, in evaluating an airline's tendency to respond to a customer's complaint, it makes more sense to restrict attention only to the initial complaining tweet posted by the customer. Therefore, in this study, we consider a tweet as an initial complaint if the customer had not communicated with the respective airline on Twitter for 8 hours before the creation of the complaining tweet under consideration.

Dependent Variable: As our dependent variable, we use a dichotomous measure that equals to one if the complaint receives a response from the airline and zero otherwise. To determine whether an airline responded to a particular complaint, we used Twitter metadata to match the user tweet with respective airline tweets, and

when the tweet was matched with one or more replies from the airline, it was considered to have received a response.

Independent Variable: The primary independent variable of interest is the customer's race, which we derive using the deep learning model described in the following sections.

Control Variables: We include a set of control variables to account for unobserved heterogeneity at the tweet level and the customer level. We also include airline fixed effects and day-of-the-week fixed effects. Table 1 explains the key variables in our empirical analysis.

Table 1. Definitions of Variables

Variable	Definition
Responded	Binary variable equal to 1 if the airline responded to the complaining tweet, 0 otherwise.
Race	Binary Variable equal to 1 if the customer is African-American, 0 otherwise
Gender	Binary Variable equal to 1 if the customer is female, 0 otherwise
Followers	Number of followers the user had, at the creation of the complaining tweet
Multiple Users Mentioned	Binary variable equal to 1 if multiple user accounts are mentioned in the complaining tweet, 0 otherwise
Complaints within the Previous Hour	Number of complaining tweets received by the airline during the hour prior to receiving the current complaining tweet
Retweets	Number of times the tweet was retweeted, before the first response from the airline (if the airline responded), or before the end of the observation period (if the airline did not respond)
Hashtag	Number of hashtags contained in the tweet
Offensive	Binary variable equal to 1 if the complaining tweet contains offensive words, 0 otherwise
URL	Binary variable equal to 1 if the complaining tweet contains web URLs, 0 otherwise
@Order	The position of the airline Twitter handle in the complaining tweet, relative to other username mentions, if any
Updates	Number of tweets ever posted by the user
Profile	Binary variable equal to 1 if the user's location, website, or profile description (i.e., Twitter bio) is publicly available, 0 otherwise
Day of Week	Categorical variable indicating the day of the week
Airline	Categorical variable indicating the airline
Cluster	Categorical variable indicating the cluster ID assigned to the complaining tweet

Behavior research in psychology shows that encountering a new individual, or facing a stimulus of human face usually activates three primitive conscious neural evaluations: race, gender, and age, which have consequential impact on the perceiver and the perceived [32, 33, 34]. Among these, race is arguably the most prominent and dominant personal trait, which is often demonstrated empirically by its omnirelevance with social, cognitive, and perceptual tasks such as attitude, biased view, stereotype, emotion and belief [35]. Hence, in this study we focus on the tweets from users (i.e., customers) having a valid profile image. As a result, users having Twitter's default profile image or inaccessible profile image URLs are not included in the analysis. Thus, we ended up with 130,023 complaints for empirical analysis.

One may be concerned that customers from different racial groups may complain about different types of problems (e.g., delays and cancellations, unprofessional employees etc.), and airline social media agents may respond to different types of problem differently. Furthermore, customers of different racial groups may also write tweets in different styles. If the tweets written by customers of different racial categories systematically differs and this difference leads to differences in response rate, then our estimation will be biased. To alleviate this concern, we use text-clustering techniques to group similar complaining tweets and introduce cluster fixed effects into our model[1].

5. Inferring Race and Gender of Twitter Users – A Deep Learning Approach

Social media platforms have become an important avenue of research to study social phenomena. However, a major limitation of the use of this social data is the lack of key user demographic indicators such as race and gender. Although social network sites usually enable their users to store demographic attributes such as gender, age, and location in their profiles, most of the time, such information is either incomplete (e.g., a user may choose not to reveal her location), or misleading (e.g., a user may choose to provide an imaginary place such as "Neverland" as her location) [36]. Therefore, in discrimination research such as this one, the foremost challenge is to effectively leverage automatic means to infer user demographics (i.e., race and gender) with a higher degree of accuracy.

Some recent studies have successfully leveraged social media data to study people's writing in a large scale. For example, recent research examined people's

[1] Text Clustering was done using the K-Means algorithm and with 40 clusters.

writings on Twitter and/or Facebook to predict their personality [37, 38, 39, 40]. Thus, in this study, we postulate that a user's social media postings may contain several important clues that describe the demographics of that user. We attempt to leverage the capabilities of deep learning, convolutional neural networks (CNN) in particular, and users' past tweets, to predict the race and gender of social media users.

5.1 Data Collection and Label Annotation

Racial and ethnic identity is complex and evaluations by others may not always match an individual's self-identification [41]. Previous studies vary significantly in their definition of race or ethnicity. For example, Pennacchiotti and Popescu [36] inferred ethnicity, but classified users as African-American or not. Chen et al. [42] claimed to infer ethnicity but used a classification scheme that uses both racial and ethnic identities. Based on the major racial categories recognized in the United States Census, in this study, race is identified in three categories: African-American, White, and Other. Here, "Hispanic or Latino" is not considered a race, since the United States Census Bureau defines it as an ethnicity rather than a race.

To obtain a more realistic dataset of users, we used Twitter's streaming API (Application Program Interface) that returns a small random sample of all public Twitter traffic. As this study intends to infer the major racial categories recognized in the U.S., and only English language is used for this study, we particularly queried for tweets that originated from the U.S. and in English. Data was streamed for two days and over 250,000 distinct users have been recognized. We further filtered this dataset for users who have a valid profile image in their Twitter account (i.e., users having Twitter's default profile image or inaccessible profile image URLs were discarded), as the ground-truth labels for the supervised classification come from profile image based human annotations. Furthermore, we retained only the users with public timelines that allows fetching their historical tweets, as the proposed classification is based on linguistic features of user-generated content. From this pool of distinct Twitter users, we randomly sampled 15,000 users for the final dataset.

Crowdsourcing human intelligence is considered an essential step in extracting demographic information encoded as images rather than text [43]. Hence, to obtain the ground-truth labels for the 15,000 random users, workers were hired from Amazon Mechanical Turk[2] (AMT) - a crowd sourcing marketplace for simple tasks that require human intelligence. For each user in the final dataset, a human intelligence task (HIT) was created on AMT that displayed the Twitter profile image of the user and the workers were asked to make an educated guess of the race of the person in the given profile image, out of the alternatives: African-American, White, Other and Can't Tell. In addition, the workers also tagged the gender of the person in the given image from the choices: Male, Female, and Can't Tell. The workers were instructed to choose the Can't Tell option only when the given image does not contain people (e.g., pets, scenery, flowers, symbols, trademarks etc.), contains multiple people, or if the image quality is not good enough to categorize (e.g., blurred images). Each HIT was labeled by three different workers choosing only those who are from the U.S. At the end of the AMT annotation process, 849 distinct workers contributed to label the 15,000 users. Only the users that received at least two agreements out of the three human annotations and those agreed upon a specific race category (i.e., users agreed upon the Can't Tell option were discarded) were reserved as valid users. Table 2 shows the statistics of labels with at least two agreed annotations.

Table 2. Label Statistics

Race		Gender	
African-American	2,841 (34.70 %)	Male	4,479 (49.93 %)
White	3,839 (46.89 %)	Female	4,492 (50.07 %)
Other	1,507 (18.41 %)		
Total	8,187	Total	8,971

Next, we used Twitter's search API to fetch the historical tweets posted by each individual user up to a maximum of 3,200 tweets. Considering the inherently noisy nature of text in tweets, first we pre-processed and cleaned the data. For example, we processed the hashtags, Internet slang words, usernames, punctuation marks, and repetitive characters in tweets before any transformation. Next, for each user, a document of text was created that contained all the unique words present in the collection of tweets posted by that user. As potential evaluation alternatives, a series of text documents was created for each user, using the last 200, 500, 1,000, and 2,000 tweets (pre-processed) each user posted. Furthermore, only the users with text documents containing at least 100 words were retained, to be used as input to the deep learning model.

5.2 Deep Learning Model

Convolutional Neural Networks and pooling architectures [44] showed a huge success in computer vison applications lately. They can identify faces,

[2] https://www.mturk.com/mturk/welcome

individuals, street signs, platypuses, and may other aspects of visual data [45]. They were introduced to the natural language processing (NLP) community in the seminal work of Collobert et al. [46] who used them for semantic-role labelling, and later by Kalchbrenner et al. [47] and Kim [48] who used them for sentence classification to predict sentiment and question-type in particular. In the context of NLP, a CNN is designed to identify indicative local clues in a large structure (irrespective of their position) and combine them to produce a fixed-size vector representation of the structure, capturing the local aspects that are most informative for the prediction task in hand [49].

Inspired by the work of Kim [48], we propose a convolutional neural network architecture to train our classifier. The model has four layers. The first layer is the input layer that contains the word embeddings of the tweet document. The second layer is the convolutional layer that contains three convolutional layers in parallel. The third layer is the max-over-time pooling layer and the final layer is the output layer. We describe each layer in detail below.

Layer 1 - Word Embeddings

The biggest jump when moving from sparse-input (e.g., bag-of-words) linear models to neural network based models is to stop representing each feature as a unique dimension (the so called one-hot representation) and representing them instead as dense vectors [49], making similar features to have similar vectors. Initializing word vectors with those obtained from an unsupervised neural language model is a quite popular method to improve performance in the absence of a large supervised training set [46, 50, 51]. In this study, we use the publicly available word2vec word embeddings published by Google[3], that were originally trained by Mikolov et al. [52] on 100 billion words of Google News. Here, each word is represented by a 300-dimensional vector. For the current study, words not present in the set of pre-trained vectors were discarded.

The first layer represents the word embeddings. Consider a user's document of tweets containing the sequence of words $X = X_1 \ldots . X_n$ each with their corresponding *k* dimensional (*k* = 300) word embedding $x_i \in \mathbb{R}^k$ representing the i^{th} word in the document. Then a document of *n* words is represented as $x_{1:n} = x_1 \oplus x_2 \oplus \ldots . . \oplus x_n$ where $\oplus$ is the concatenation operator. This vector holds the information in the entire document. In the next layer, a convolutional operation or a filter is applied to the word embeddings in the first layer, on a sliding window basis.

Layer 2 – Convolutional Layer

The main idea behind a convolution for language tasks is to apply a non-linear (learned) function over each instantiation of a *h*-word sliding window over the text (Goldberg 2016). A 1d convolution layer of width *h* works by moving a sliding window of *h* words over the document and applying the same filter $w \in \mathbb{R}^{hk}$ to each window in the sequence $x_{i:i+h-1}$ to produce a new feature. For example, if the user document contained the text "flights delayed and the baby next to me crying", and with a sliding window of size two, the filter will be applied to the window of first two words "flights delayed", then to the next two words "delayed and", and then to the next two words "and the", and so on. The filter basically transforms a window of *h* words into a vector that captures important properties of the words in the window. Applying the convolution operation (i.e., filter) on a window of words $x_{i:i+h-1}$ generates a feature c_i, $c_i = f(w.x_{i:i+h-1} + b)$ where $b \in \mathbb{R}$ is a bias term and f is a non-linear activation function that is applied element-wise. Both *w* and *b* are parameters to be estimated. Thus, the convolution operation applies a linear transformation to the input and then a non-linear transformation using the activation function. In this study, we use the Leaky Rectified Linear Units (Leaky RELU) [53] as the activation function. Leaky RELU is defined as $f(x) = max(0.01x, x)$. The same filter is applied to each possible window of words in the document $(x_{i:h}, x_{2:h+1}, \ldots, x_{n-h+1:n})$ producing a feature map *c*, $c = [c_1, c_2, \ldots, c_{n-h+1}]$ where $c \in \mathbb{R}^{n-h+1}$. We use three different window sizes (*h*= 3, 4, 5) to include tri-grams, 4-grams, and 5-grams.

Layer 3 - Max-over-time Pooling:

The pooling layer applies a max-over-time pooling operation [46] over the feature map created in the layer-2 and take the maximum value $\hat{c} = \max\{c\}$ as the feature corresponding to this filter. The resulting vector $\hat{c}$ is a representation of the entire document in which each dimension reflects the most salient information with respect to the prediction task.

Layer 1-3 explained above extract one feature from one filter. To obtain multiple features, we repeat this process with multiple filters and varying window sizes. The pooling operation creates one feature from each filter which will be concatenated and passed to the next layer. In our context, we use a total of 100 filters for

[3] https://code.google.com/archive/p/word2vec/

each window size (h = 3, 4, 5). These features form the penultimate layer and are passed to a fully connected softmax layer (Layer 4 – Output Layer) whose output is the probability distribution over labels. The deep learning model was trained by back propagation and gradient-based optimization using the adaptive learning rate algorithm Adam as the updating rule. As our primary objective of using this model for the study is to predict whether a given user is African-American or not, we built a binary classifier to predict race of social media users. For the race classifier, we use a balanced dataset of 5,000 users with 85% assigned for the training set and 15% for the test set.

Similarly, a separate classifier to categorize users based on their gender was also developed, using 7,000 users. The same training and test data separation rule of the race classifier has also been used. The prediction performances of the classes of interest in our CNN model are reported in Table 3.

Table 3. CNN Model – Prediction Performance

Class	Precision (%)	Recall (%)	F1 Score (%)
African-American	81.07	98.5	88.94
Gender - Female	92.51	88.29	90.35

In general, both the race classifier and the gender classifier achieve good precision, recall and F1 scores. Furthermore, the race classifier achieves an overall prediction accuracy and F1 score of approximately 88% (average across all classes), while the overall prediction accuracy and the F1 score for the gender classifier is also about 91%. These outperform or match most of the existing text-based user classification methods. Next, we used these race and gender classifiers to derive the race and the gender of social media users in our main dataset.

6. Econometric Analysis and Results

We assume that for airline k, the perceived value of responding to the complaining tweet i created by customer j is Y_{ijk}^* where

$$Y_{ijk}^* = \beta_0 + C_{ij}\beta_1 + T_i\beta_2 + \alpha_k + D_t + \varepsilon_{ijk}$$

Here C_{ij} refers to the vector of observable characteristics of customer j at the creation of complaining tweet i and T_i refers to the vector of observable characteristics related to complaining tweet i. D_t is the day of week fixed effect, and α_k is the airline fixed effect. We assume the error term ε_{ijk} follows a logistic distribution.

The airline chooses to respond to the tweet if the perceived value of responding $Y_{ijk}^* > 0$. To test our hypothesis, we estimate this model and present the results in Table 4.

Table 4. Estimation Results

Variable	Logit Coefficient
African-American (Base: Other)	**-0.1282*****(0.0322)
Gender - Female (Base: Male)	-0.0279*(0.0143)
Log of Followers	0.0904***(0.0052)
Multiple Users Mentioned	-0.7711***(0.0299)
Log of Complaints in the last hour	-0.4340***(0.0081)
Log of Retweets	-0.6521***(0.0305)
Hashtag	0.0038(0.0083)
Offensive	-0.4069***(0.0490)
URL	-0.6548***(0.0236)
@Order	-0.6854***(0.0250)
Log of Updates	-0.1095***(0.0051)
Profile	0.0568**(0.0243)
Constant	3.8093***(0.1067)
Observations	130,023

From Table 4, the variable *African-American* is negative and statistically significant ($p<0.01$). In particular, being an African-American customer decreases the odds of getting a response by the airline by 12.03%, compared to Non-African-Americans. Therefore, our findings provide evidence to support the racial discrimination hypothesis. Another interesting finding is that the variable *Gender* shows negative and statistically significant ($p<0.1$) effects on receiving a response from an airline. In other words, being a female customer decreases the odds of hearing back from the airline by 2.75%, compared with male customers.

7. Discussion and Conclusion

Our empirical results show very strong evidence that airlines are less likely to respond to complaints sent to it by African-American customers, than to the customers who are not African-American, suggesting potential racial discrimination in social media customer service. Furthermore, we find that female customers are also less likely to receive a response from an airline compared to male customers, although the effect is marginally significant.

The contribution of this study to the field of information systems research is twofold. First, to the best of our knowledge, this is the first study to empirically analyze the racial discrimination phenomenon in the context of social media customer

service. Second, we propose a methodology to effectively leverage deep learning, convolutional neural networks in particular, and social media data to predict latent user attributes (i.e., race and gender) of social media users. This can be particularly useful for social science researchers, as social science research often relies on demographic attributes to characterize and group users or survey participants. Our findings have important implications, especially for companies that are trying to harness the power of social media to provide customer service. The empirical test of our hypothesis provides evidence that suggests potential racial discrimination in social media customer service. Given its controversial nature, it is of vital importance that companies carefully examine the drivers of such a practice in their social media customer service and define appropriate action accordingly. Even if the driving force of this practice could be implicit/unconscious bias, that would not absolve one from the responsibility to be aware of the morally discriminatory aspects of behavior resulting from these biases [54]. This calls for the need to offer mandatory racial discrimination prevention training programs for the social media teams in addition to having anti-discriminatory policies in place.

8. References

[1] Edelman, B., Luca, M., and Svirsky, D. 2017. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment," *American Economic Journal: Applied Economics* (9:2), pp.1-22.

[2] Gabbidon, S. L. 2003. "Racial Profiling by Store Clerks and Security Personnel in Retail Establishments: An Exploration of "Shopping While Black", *Journal of Contemporary Criminal Justice* (19:3), pp. 345-364.

[3] Borjas, G. J., and Bronars, S. G. 1989. "Consumer Discrimination and Self-employment," *Journal of Political Economy* (97:3), 581-605.

[4] Schreer, G. E., Smith, S., and Thomas, K. 2009. "Shopping while Black: Examining Racial Discrimination in a Retail Setting," *Journal of Applied Social Psychology* (39:6), pp.1432-1444.

[5] Harris, A. M. G., Henderson, G. R., and Williams, J. D. 2005. "Courting Customers: Assessing Consumer Racial Profiling and Other Marketplace Discrimination," *Journal of Public Policy & Marketing* (24:1), pp.163-171.

[6] Ayres, I., and Siegelman, P. 1995. "Race and Gender Discrimination in Bargaining for a New Car," The American Economic Review (85:3), pp. 304-321.

[7] Morton, F. S., Zettelmeyer, F., and Silva-Risso, J. 2003. "Consumer Information and Discrimination: Does the Internet Affect the Pricing of New Cars to Women and Minorities?," *Quantitative Marketing and Economics* (1:1), pp. 65-92.

[8] Acquisti, A. and Fong, C. M. 2015. "An Experiment in Hiring Discrimination Via Online Social Networks," Available at SSRN: https://ssrn.com/abstract=2031979

[9] Sánchez Abril, P., Levin, A., and Del Riego, A. 2012. "Blurred Boundaries: Social Media Privacy and the Twenty-First-Century Employee," *American Business Law Journal* (49:1), pp. 63-124

[10] Ge, Y., Knittel, C. R., MacKenzie, D., and Zoepf, S. 2016. "Racial and gender discrimination in transportation network companies," Working Paper (No. w22776), National Bureau of Economic Research.

[11] Laouénan, M. and Rathelot, R. 2017. "Ethnic Discrimination on an Online Marketplace of Vacation Rentals," Working Paper Series.

[12] Bartoš, V., Bauer, M., Chytilová, J., and Matějka, F. 2016. "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition," American Economic Review (106:6), pp.437-75.

[13] Cherry, R. 1976. "Racial Thought and the Early Economics Profession," *Review of Social Economy* (34:2), pp.147-162.

[14] Reich, M. 1981. *Racial Inequality: A Political-economic Analysis*. Princeton University Press.

[15] Becker, G.S., 1957. The Economics of Discrimination. University of Chicago press.

[16] Guryan, J. and Charles, K.K., 2013. "Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots," *The Economic Journal* (123:572).

[17] Phelps, E.S., 1972. "The Statistical Theory of Discrimination," *American Economic Review* (62:4), pp.659-661.

[18] Arrow, K., 1973. "The Theory of Discrimination," *Discrimination in Labor Markets* (3:10), pp.3-33.

[19] Bertrand, M., and Duflo, E. 2017. "Field Experiments on Discrimination. Handbook of Economic Field Experiments," 1, pp. 309-393.

[20] Tajfel, H. 1970. "Experiments in Intergroup Discrimination," *Scientific American* (223:5), pp. 96-103.

[21] Tajfel, H., and Turner, J. C. 1979. "An Integrative Theory of Intergroup Conflict," *The Social Psychology of Intergroup Relations* (33:47), 74.

[22] Banaji, M. R., and Greenwald, A. G. 1995. "Implicit Gender Stereotyping in Judgments of Fame," Journal of Personality and Social Psychology (68:2), pp.181-198.

[23] Bertrand, M., Chugh, D., and Mullainathan, S. 2005. "Implicit Discrimination," American Economic Review (95:2), pp.94-98.

[24] Greenwald, A. G., and Banaji, M. R. 1995. "Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes," *Psychological Review* (102:1), pp.4-27.

[25] Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., and Phelps, E. A. 2009. "A neural mechanism of first impressions," *Nature neuroscience* (12:4), pp. 508-514.

[26] Hugenberg, K., Young, S. G., Bernstein, M. J., and Sacco, D. F. 2010. "The categorization-individuation model: an integrative account of the other-race recognition deficit," *Psychological Review* (117:4), pp. 1168.

[27] Bigler, R. S., and Liben, L. S. 2006. "A Developmental Intergroup Theory of Social Stereotypes and Prejudice," *Advances in Child Development and Behavior* (34), pp. 39-89.

[28] Jones, C. R., and Fazio, R. H. 2010. "Person categorization and automatic racial stereotyping effects on weapon identification," *Personality and Social Psychology Bulletin* (36:8), pp. 1073-1085.

[29] Aigner, D. J., and Cain, G. G. 1977. "Statistical Theories of Discrimination in Labor Markets," ILR Review (30:2), pp.175-187.

[30] Yinger, J. 1998. "Evidence on Discrimination in Consumer Markets," *The Journal of Economic Perspectives* (12:2), pp. 23-40.

[31] Fix, M. and Struyk, R. J. 1993. *Clear and Convincing Evidence: Measurement of Discrimination in America*, Urban Institute Press, Washington, D.C.

[32] Bruce, V., and Young, A. 1986. "Understanding Face Recognition," *British Journal of Psychology* (77:3), pp. 305-327.

[33] Calder, A. (Ed.). 2011. Oxford handbook of face perception. Oxford University Press.

[34] Calder, A. J., and Young, A. W. 2005. "Understanding the recognition of facial identity and facial expression," *Nature Reviews Neuroscience* (6:8), pp.641-651.

[35] Fu, S., He, H., and Hou, Z. G. 2014. "Learning Race from Face: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (36:12), pp.2483-2509.

[36] Pennacchiotti, M., and Popescu, A. M. 2011. "A Machine Learning Approach to Twitter User Classification," In *Proceedings of the Fourth ICWSM 2011*, pp. 281-288.

[37] Golbeck, J., Robles, C., Edmondson, M., and Turner, K. 2011a. Predicting Personality from Twitter. In *IEEE Third International Conference on Social Computing (SocialCom)*, IEEE Computer Society Press: Washington, DC, pp. 149-156.

[38] Golbeck, J., Robles, C., and Turner, K. 2011b. Predicting Personality with Social Media. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, ACM, May 2011, 253-262.

[39] Gou, L., Zhou, M. X., and Yang, H. 2014. Know Me and Share Me: Understanding Automatically Discovered Personality Traits from Social Media and User Sharing Preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York: ACM Press, pp. 955-964.

[40] Sumner, C., Byers, A., Boochever, R., and Park, G. J. 2012. Predicting dark triad personality traits from Twitter usage and a linguistic analysis of tweets. In *Proceedings of the 2012 11th International Conference on Machine Learning and Applications (ICMLA)*, Washington, DC: IEEE Computer Society Press, pp. 386-393.

[41] Cesare, N., Grant, C., and Nsoesie, E. O. 2017. "Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices," *arXiv* preprint *arXiv*:1702.01807.

[42] Chen, X., Wang, Y., Agichtein, E., and Wang, F. 2015. "A Comparative Study of Demographic Attribute Inference in Twitter," In *Proceedings of the ICWSM* 2015, pp. 590-593.

[43] McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., and Spiro, E. S. 2017. "Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing," *Sociological Methods & Research* (46:3), pp. 390-421.

[44] LeCun, Y., and Bengio, Y. 1995. "Convolutional Networks for Images, Speech, and Time Series," *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed. Cambridge, MA: MIT Press, 1995, pp. 255–258.

[45] Patterson, J., and Gibson, A. 2017. Deep Learning for Practitioners, Sebastopol CA: O'Reilly Media Inc.

[46] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuglu, K., Kuksa, P. 2011. "Natural Language Processing (Almost) from Scratch," *Journal of Machine Learning Research* (12), pp. 2493–2537.

[47] Kalchbrenner, N., Grefenstette, E., Blunsom, P., Kartsaklis, D., Kalchbrenner, N., Sadrzadeh, M., and Blunsom, P. A. 2014. "Convolutional Neural Network for Modelling Sentences," In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 212-217.

[48] Kim, Y. 2014, "Convolutional Neural Networks for Sentence Classification," In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2014, October 25-29, 2014, Doha, Qatar, pp. 1746–1751.

[49] Goldberg, Y. 2016, "A Primer on Neural Network Models for Natural Language Processing," *Journal of Artificial Intelligence Research* (57), pp. 345–420.

[50] Socher, R., Pennington, J., Huang, E., Ng, A., Manning, C., 2011. "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions," In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, pp. 151-161.

[51] Iyyer, M., Enns, P., Boyd-Graber, J., Resnik, P. 2014. "Political Ideology Detection Using Recursive Neural Networks," In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 1113-1122.

[52] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. 2013. "Distributed Representations of Words and Phrases and their Compositionality," *Advances in Neural Information Processing Systems*, pp. 3111-3119.

[53] Maas, A. L., Hannun, A. Y., and Ng, A. Y. 2013. "Rectifier Nonlinearities Improve Neural Network Acoustic Models," In *Proceedings of International Conference on Machine Learning 2013*.

[54] Holroyd, J. (2015). "Implicit Bias, Awareness, and Imperfect Cognitions," *Consciousness and Cognition* (33), pp. 511–523.